Human Activity Recognition from Smartphone Accelerometer Data

John Gaber
SYDE 522

University of Waterloo

***Abstract*—This project investigates the use of machine learning methods to classify physical activities-walking, running, and sitting-based on smartphone accelerometer data. The goal is to evaluate and compare the effectiveness of classical algorithms and deep learning models in recognizing these activities from raw sensor input.**

We use the WISDM dataset, which includes tri-axial acceleration data sampled at 20Hz from 51 participants. Input signals are divided into fixed-length windows and processed through multiple algorithms: k-Nearest Neighbors, Support Vector Machines, Multilayer Perceptrons, Convolutional Neural Networks, and both unidirectional and bidirectional Long Short-Term Memory networks. For each algorithm, we conduct a parameter sweep and run each configuration five times to evaluate stability.

Results show that CNNs achieve the highest accuracy among deep models, with a peak F1 score of 0.796. SVMs slightly outperform deep networks, reaching an F1 score of 0.807. Recurrent models underperform, likely due to the short, fixed-length input windows.

These findings suggest that models sensitive to local signal features or capable of strong boundary-based separation are most effective for this task. The results offer guidance for real-time activity recognition systems that prioritize both accuracy and efficiency.

I. Introduction and Background

In this project, we use AI to classify accelerometer data from a smartphone to determine whether the person carrying the smartphone is performing a certain action. This involves training models to recognize patterns in tri-axial acceleration signals corresponding to these physical activities.

Human Activity Recognition (HAR) from smartphone sensors is a widely useful task with applications in healthcare, fitness, smart homes, security, and more [1]. For example, a mobile app could automatically track a user's exercise or monitor an elderly person's activity level. Continuous activity monitoring provides personalized health insights, especially valuable for older adults [1]. Compared to camera-based monitoring, smartphone sensors offer a privacy-preserving, infrastructure-free approach [2], since most people already carry smartphones.

Research in this area has been ongoing for over a decade. Early approaches relied on manually extracted features and traditional classifiers like SVM, kNN, or decision trees. For example, a combination of these methods achieved about 87% accuracy on accelerometer-based activity recognition [3]. Deep learning has improved performance, particularly with CNN and LSTM models. A bidirectional LSTM model recently reached 97.5% accuracy [4], while CNN-based ensemble models also exceed 90% accuracy [2]. Depending on the application, different metrics may be found.

Identify applicable funding agency here. If none, delete this.

A. Dataset

We use the WISDM dataset [5], which consists of raw tri-axial accelerometer data collected from 36 participants using smartphone sensors. Participants performed a set of daily activities including walking, jogging, walking upstairs, walking downstairs, sitting, and standing. Accelerometer signals were recorded at a sampling rate of 20 Hz.

The continuous signals are segmented into fixed-length temporal windows, each containing acceleration measurements along the X, Y, and Z axes. In the original study, windows of approximately 10 seconds were used, although alternative window lengths may be employed depending on the modeling approach. Input representations may consist of raw accelerometer values or derived features such as statistical moments (e.g., mean, variance) and signal energy.

The target outputs are categorical activity labels (e.g., walking, jogging, sitting). To evaluate generalization to unseen individuals, we partition the data into training, validation, and testing subsets (70%, 15%, and 15%, respectively).

B. Algorithms and Parameters

We evaluate a range of classification algorithms, including both traditional machine learning models and deep learning models, to identify which approaches are most effective for recognizing physical activities from smartphone sensor data.

k-Nearest Neighbors (kNN): A non-parametric algorithm that classifies each sample based on the majority label of its k closest neighbors in the training set, according to a distance metric. We experiment with $k = 3, 5, 7$ and compare the Euclidean and Manhattan distance measures.

Support Vector Machine (SVM): A classifier that seeks an optimal boundary between activity classes in a high-dimensional space. We use the radial basis function (RBF) kernel to allow for non-linear decision boundaries, and tune the regularization parameter C and kernel coefficient γ to balance model complexity and performance.

Multilayer Perceptron (MLP): A fully connected feedforward neural network trained on raw data. We vary the number of hidden units (e.g., 64, 128, 256). MLPs rely on learned transformations of input features but do not explicitly model time.

Convolutional Neural Network (CNN): A neural network architecture well-suited to extracting spatial or local patterns. Here, 1D convolutional layers scan across the time axis of the input sequences, detecting activity-specific motion patterns. We evaluate different kernel sizes (e.g., 3, 5, 7) and apply dropout regularization to prevent overfitting.

Long Short-Term Memory (LSTM): A type of recurrent neural network that is capable of modeling sequential dependencies over time. LSTMs maintain memory cells that capture context from previous time steps. We experiment with different hidden state sizes and also test bidirectional LSTMs, which process sequences in both forward and backward directions.

For all models, we report validation accuracy and weighted F1-score as the primary performance metrics. Confusion matrices are also used to analyze class-specific performance. Confidence intervals are recorded to show statistical variability.

II. Eperimental setup and Results

A. Evaluation Metrics

To evaluate the performance of the models, we use two primary metrics: classification accuracy and the weighted F1-score. Accuracy measures the proportion of correct predictions, while the F1-score accounts for both precision and recall, making it more informative when class distributions are imbalanced. Each experiment was run five times to account for performance variability.

B. Baseline Classifiers

We first evaluated traditional machine learning classifiers using statistical features derived from the raw accelerometer windows. For each 5-second window, we computed the mean, standard deviation, minimum, and maximum of each axis. These features were then used to train KNN and SVM classifiers.

For kNN, we tested $k \in \{3, 5, 7\}$ with both Euclidean and Manhattan distances. For SVM, we used an RBF kernel and teseted the regularization parameter $C \in \{0.1, 1, 10\}$ and the kernel coefficient at $\gamma \in \{\texttt{scale}, \texttt{auto}\}$.

Results showed that SVM with $C = 1$ and $\gamma = \texttt{scale}$ achieved the highest performance among classical models, with 81.2% accuracy and a weighted F1-score of 0.807. The best kNN configuration (k=7, Euclidean) achieved 74.8% accuracy and 0.757 F1-score.

C. Deep Learning Models

We next evaluated neural network architectures trained directly on the raw windowed data. All deep learning models were implemented using TensorFlow and trained using categorical cross-entropy loss and the Adam optimizer. Input data was normalized across all time steps and axes.

Multilayer Perceptron: We flattened the windowed data and trained fully connected networks with 2 hidden layers. We varied the hidden size (64, 128, 256 neurons). The best result was obtained with 128 hidden units, yielding 72.0% accuracy and 0.735 F1-score.

Convolutional Neural Network: The CNN consisted of one convolutional layer followed by max pooling, dropout, and dense layers. We varied the kernel size ($k = 3, 5, 7$). The best performing configuration used $k = 3$ with a 0.804 accuracy and 0.796 F1-score.

Long Short-Term Memory: We trained unidirectional LSTMs with hidden sizes of 32, 64, and 128. The best accuracy was observed with 32 hidden units (70.2%), though the F1-score peaked with 64 hidden units at 0.621.

Bidirectional LSTM : Similar to LSTM, we tested hidden sizes of 32, 64, and 128. The model with 32 hidden units achieved the best performance at 69.7% accuracy and 0.605 F1-score.

D. Summary of Results

Table I summarizes the best configuration for each model based on validation accuracy. Confidence intervals are shown for deep learning models where applicable. Figure 1 and Figure 2 visualize the comparative performance of all models.

TABLE I
Best Validation Performance per Model

Model	Accuracy	F1-score
SVM (C=1)	0.812	0.807
kNN (k=7, Euclidean)	0.748	0.757
MLP (hidden=128)	0.720 (0.694–0.745)	0.735 (0.709–0.761)
CNN (kernel=3)	0.804 (0.768–0.841)	0.796 (0.768–0.824)
LSTM (hidden=64)	0.699 (0.676–0.721)	0.621 (0.595–0.647)
BiLSTM (hidden=32)	0.697 (0.692–0.703)	0.605 (0.575–0.636)

E. Confusion Matrices and Class-wise Analysis

To better understand model performance across activities, we examined confusion matrices for each model. CNN and SVM exhibited the highest per-class accuracy, with CNN making the fewest misclassifications overall (Fig. 3, Fig. 4). MLP also performed well, though it showed some confusion between walking and running (Fig. 5). LSTM-based models, especially BiLSTM, commonly made the mistake of labeling walking up or down stairs as simply walking (Fig. 6, Fig. 7). The kNN classifier was generally accurate but less robust to activity transitions compared to the deep learning models (Fig. 8).

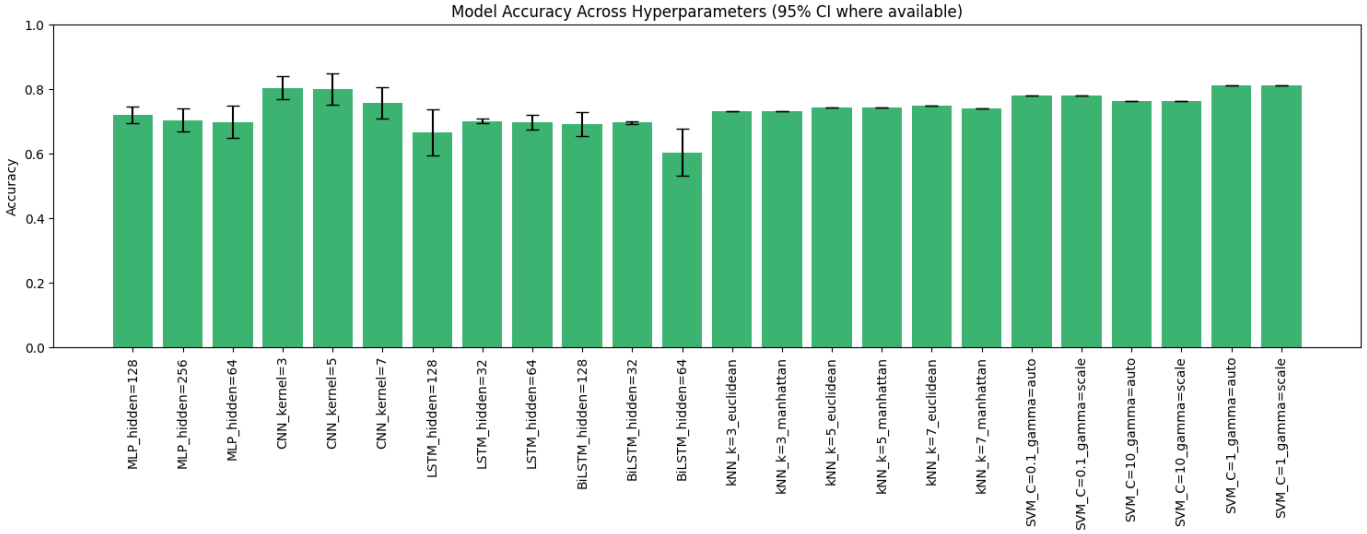

Fig. 1. Validation accuracy across all models and parameter configurations. Models include classical algorithms and deep learning models.

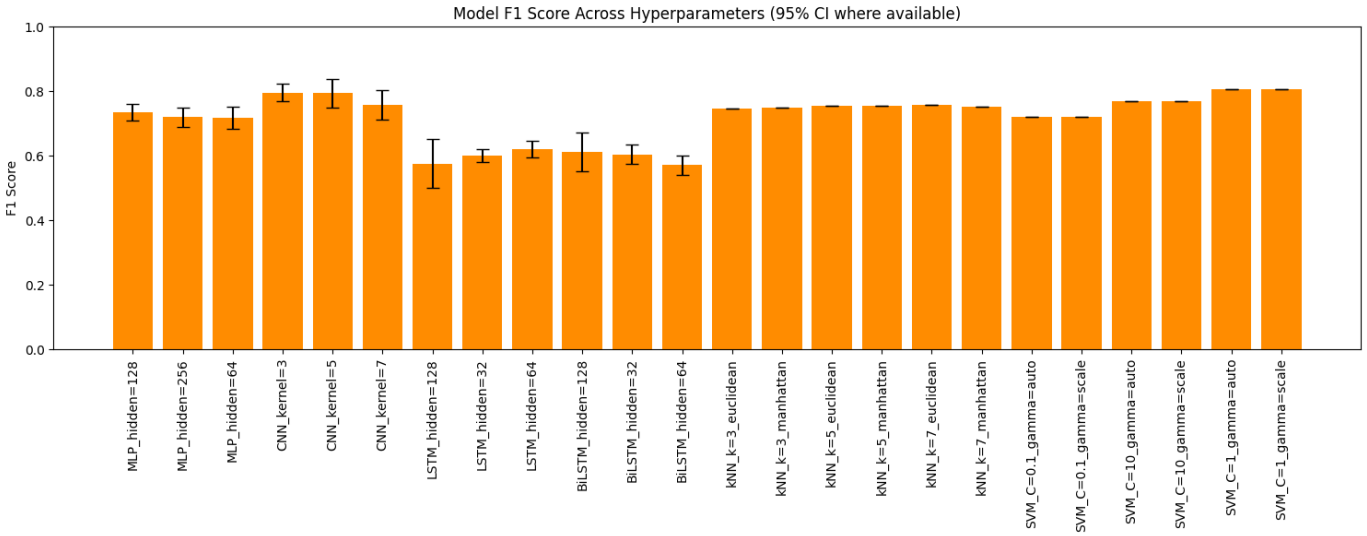

Fig. 2. F1 scores across all models and parameter configurations. These scores reflect the balance between precision and recall in activity classification.

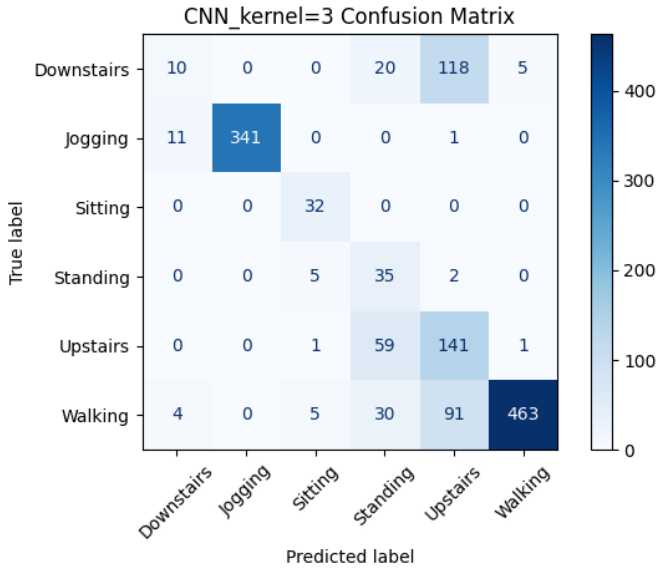

Fig. 3. Confusion matrix for CNN.

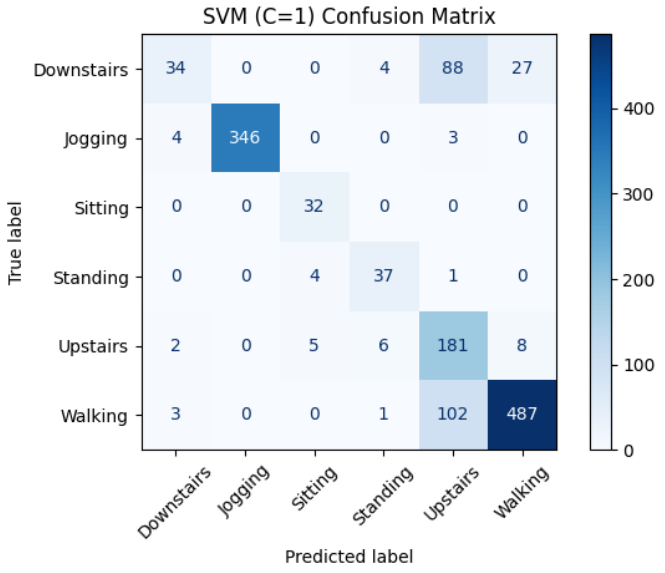

Fig. 4. Confusion matrix for SVM.

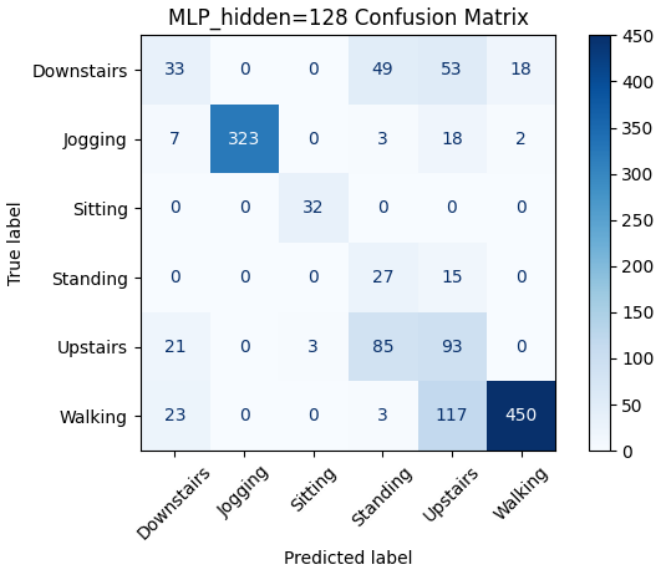

Fig. 5. Confusion matrix for MLP.

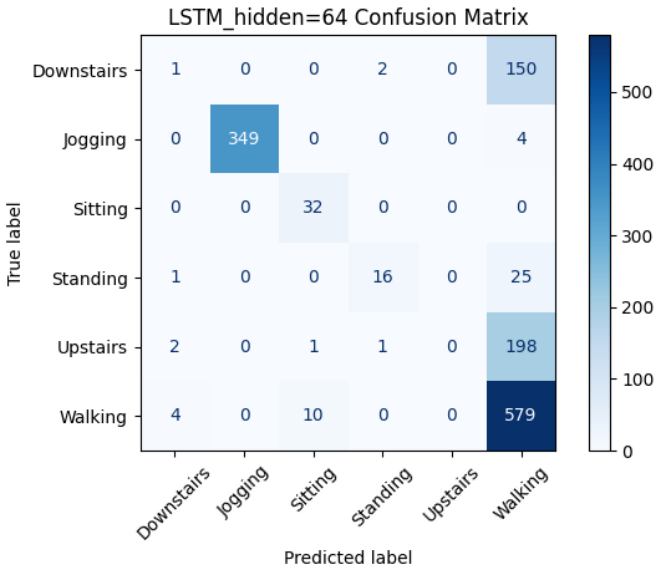

Fig. 6. Confusion matrix for LSTM.

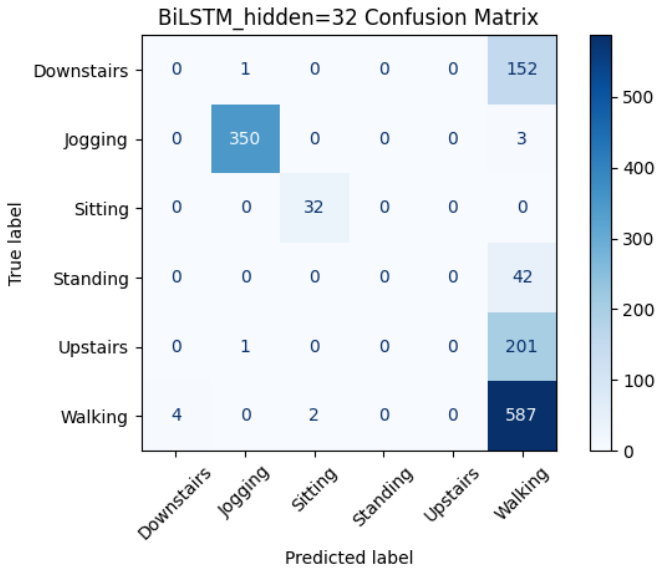

Fig. 7. Confusion matrix for BiLSTM.

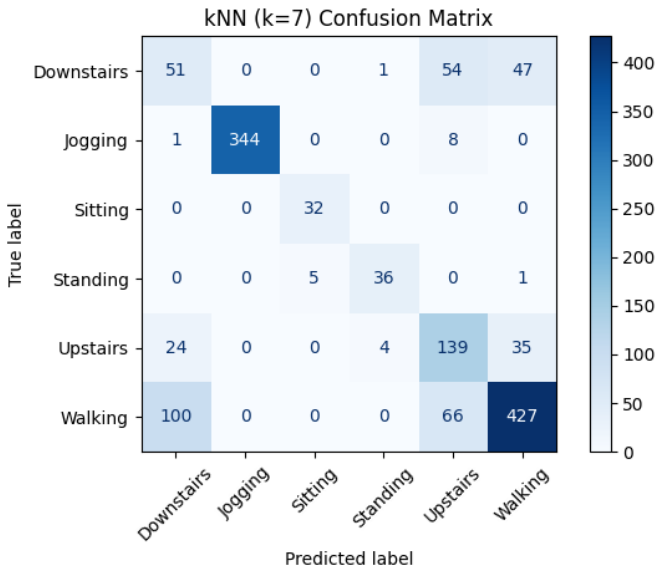

Fig. 8. Confusion matrix for kNN.

III. DISCUSSION

The strongest-performing models were the convolutional neural network with a kernel size of 3 and the support vector machine with a regularization parameter of 1 and gamma set to scale. Both models consistently achieved high accuracy and F1 scores, indicating that they were capable of distinguishing between activities based on short windows of acceleration data.

The convolutional network's success can be attributed to its ability to extract local temporal patterns from the sensor signal without requiring manually engineered features. Meanwhile, the support vector machine demonstrated that classical feature-based methods still offer competitive performance.

The multilayer perceptron performed reliably across different hidden layer sizes, but it was limited by its inability to model sequential structure. Its results suggest that dense architectures are capable of learning broad trends in the data but may struggle with finer temporal distinctions.

The recurrent models exhibited notably lower performance. Despite being designed for sequential data, both the unidirectional and bidirectional LSTM models failed to outperform simpler baselines. The lack of clear improvement suggset that, under the constraints of this dataset and task, their sequence modeling capacity was underutilized. It is likely that the limited window length and the simplicity of the activity categories reduced the benefit of temporal recurrence.

Misclassification trends further emphasize these findings. The confusion matrices show that stair related activities were often labeled as walking, particularly by the LSTM and BiLSTM models. This overlap likely stems from the similarity in motion patterns, as all involve rhythmic leg movement and similar acceleration profiles. In this context, models like CNNs and SVMs perform better because they can either focus on fine-grained local features in the signal or create clear decision boundaries between overlapping classes. This helps explain why these models misclassify less frequently than LSTM-based architectures, which may be less suited to short, windowed sequences where long-term temporal dependencies are limited

Overall, results indicate that task-specific architectural choices, rather than complexity alone, determine performance in human activity recognition from smartphone accelerometer data.

IV. CONCLUSION

This study compared classical and deep learning models for activity recognition from smartphone accelerometer data. Using the WISDM dataset, we benchmarked models including kNN, SVM, MLP, CNN, LSTM, and BiLSTM. Results showed that both CNN and SVM achieved high performance, with CNN slightly outperforming SVM in accuracy, while SVM achieved the highest F1-score overall.

The value of both traditional and deep learning approaches in Human Activity Recognition was demonstrated by the results, particularly CNNs are a strong choice when raw data is available and computational resources permit, while SVMs remain a robust option for low-latency or resource-constrained scenarios.

Future work may explore sensor fusion (combining accelerometer with other sensors like a gyroscope), larger datasets, and real-time deployment considerations. Additionally, augmenting the dataset with more complex activities or dynamic environments could further test model generalization.

REFERENCES

[1] Z. Zhong and B. Liu, "Efficient human activity recognition using machine learning and wearable sensor data," *Applied Sciences*, vol. 15, no. 8, p. 4075, Apr. 2025.

[2] S. Kundu, M. Mallik, J. Saha, and C. Chowdhury, "Smartphone based human activity recognition irrespective of usage behavior using deep learning technique," *International Journal of Information Technology*, vol. 17, no. 1, pp. 69–85, Sep. 2025.

[3] L. A. Silva Zendron, P. J. Coelho, C. Soares, I. Pereira, and I. M. Pires, "Enhancing human activity recognition with machine learning: Insights from smartphone accelerometer and magnetometer data," *PeerJ Computer Science*, vol. 11, Sep. 2025.

[4] L. Xiao, K. Luo, J. Liu, and A. Foroughi, "A hybrid deep approach to recognizing student activity and monitoring health physique based on accelerometer data from smartphones," *Scientific Reports*, vol. 14, no. 1, Jun. 2024.

[5] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity recognition using cell phone accelerometers," *SIGKDD Explorations*, vol. 12, no. 2, pp. 74–82, 2011.